Un système de dictionnaire de mots simples du coréen

Sébastien Paumier, Jee-Sun Nam

Sébastien Paumier[1] & Jeesun Nam[2]

[1] LIGM, Université Paris-Est Marne-la-Vallée, [2] HUFS

[1] paumier@univ-mlv.fr [2] namjs@hufs.ac.kr

UN SYSTÈME DE DICTIONNAIRE DE MOTS SIMPLES DU CORÉEN

Résumé

Les lexiques des langues agglutinantes ne se prêtent pas à une représentation par liste d'entrées, car la combinatoire des morphèmes est si grande qu'elle produirait un dictionnaire gigantesque. Une façon de contourner ce problème est de représenter de tels lexiques directement sous une forme factorisée, en particulier à l'aide d'automates. Dans cet article, nous présentons une description d'un tel système pour le coréen. Ce système est pleinement opérationnel, et a déjà fait l'objet d'adaptation pour d'autres langues agglutinantes.

Mots-clés: coréen, dictionnaire électronique, langue agglutinante, TAL, automates

1. Introduction

Les analyseurs morphologiques se divisent schématiquement en deux grandes catégories: les systèmes à base de règles de calcul, utilisant ou non de l'apprentissage automatique (Koskenniemi 1984, Beesley & Karttunen 2003, Han & Palmer 2005) et ceux reposant sur des lexiques construits manuellement par des linguistes (Gross 1989, Courtois 1990, Silberztein 1993). Les premiers offrent une économie de main d'œuvre lors de l'adaptation à une nouvelle langue et une certaine tolérance à l'erreur. Les seconds garantissent une meilleure précision. Le système que nous proposons s'inscrit dans cette deuxième catégorie. L'approche classique consiste à produire un lexique sous forme de liste d'entrées et à le transformer ensuite en un format plus propice à une exploitation logicielle, le plus souvent sous forme d'automate, ce formalisme étant particulièrement adapté à cette tâche (Revuz 1991, Roche & Schabès 1997).

Toutefois, il n'est pas possible d'utiliser cette méthode pour des langues agglutinantes comme le coréen, car la combinatoire des morphèmes est telle qu'un dictionnaire sous forme de liste occuperait une taille gigantesque. Il est donc nécessaire de construire directement le lexique sous la forme d'un automate qui factorise les morphèmes et évite l'explosion combinatoire. De premiers prototypes d'un tel système ont été proposés pour le coréen par (Lee 1997) et (Huh 2005), mais des problèmes d'architecture, de formats de fichiers et de maintenance les rendaient difficiles à manipuler, non seulement par les utilisateurs finaux de ces analyseurs, mais, ce qui est plus problématique, également par les linguistes chargés de produire les données. En effet, si la description d'un lexique sous forme d'une liste d'entrées est aisément manipulable par un linguiste, la nécessité de gérer l'agglutination introduit une complexification du formalisme de description pouvant considérablement dégrader son utilisabilité réelle si la tâche du créateur de ressources en devient trop compliquée.

Nous décrivons dans cet article une nouvelle version de ce système de dictionnaire, beaucoup plus simple d'utilisation, et généralisable aux autres langues agglutinantes. Nous avons conservé le principe d'une description du dictionnaire directement sous forme d'automates, mais en déplaçant au maximum la complexité qui se trouvait jusque-là dans les données elles-mêmes vers les programmes chargés de les manipuler, réduisant ainsi au minimum les efforts d'adaptation demandés aux linguistes produisant les dictionnaires, notamment en terme de lisibilité et de maintenabilité des données, critères toujours cruciaux dès lors qu'il y a intervention humaine. Ce système a été intégré au logiciel libre de traitement de corpus Unitex (Paumier 2010).

2. Architecture générale du système

La majeure partie des mots simples en coréen est constituée d'une racine à laquelle vient se combiner une série de postpositions. Ainsi dans le DECO (Dictionnaire Electronique du COréen), les quatre catégories Nom (NS), Verbe (VS), Adjectif (AS) et aDverbe (DS) sont enregistrées avec les codes flexionnels indiquant les classes des postpositions attachables, alors que la catégorie Determinant (TS) ne demandant aucune série de postpositions est intégrée sans le code flexionnel (Nam 2002, 2003, 2007). Les tokens en coréen dits Eojeol sont une unité plus grande qu'un mot en français, ce qui cause une complexité sérieuse de l'analyse morphologique et une ambiguïté plus grave qu'en français. De plus, dans les cas des verbes et des adjectifs, la racine peut subir des variations morphologiques qui conduisent à l'obtention d'une ou plusieurs variantes, chacune pouvant se combiner avec une certaine classe de postpositions. Dans la discussion suivante, nous allons détailler les différentes composantes du système avec le cas des verbes.

2.1 Génération des variantes des racines

La génération des variantes des racines suit exactement la même logique que la procédure de flexion automatique utilisée pour les langues non-agglutinantes (Silberztein 1999). Le principe est de recenser les formes canoniques en leur associant des codes qui décrivent leur paradigme flexionnel. Ces paradigmes sont décrits sous la forme d'automates décrivant des opérateurs à appliquer sur la forme canonique pour obtenir les formes fléchies, à l'aide d'un mécanisme de pile.

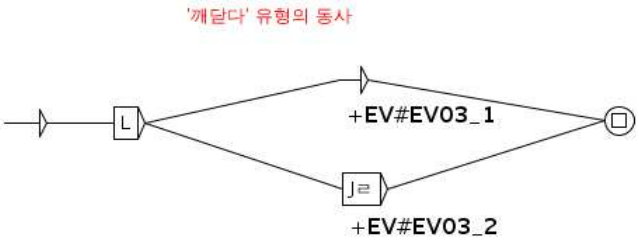

Figure 1: graphe générant les variantes des racines de la classe VS03

Par exemple, le graphe de la figure 1 permet d'obtenir deux variantes à partir d'une racine donnée. Le `L` commun aux deux chemins indique qu'on doit retirer un caractère syllabique Hangul. Le chemin du haut indique qu'on obtient, sans autre modification, une nouvelle racine dotée du code `EV#EV03_1`. Dans le chemin du bas, la séquence `J ㄹ` indique qu'on doit retirer une lettre Jamo et ensuite ajouter la lettre `ㄹ`. La racine ainsi obtenue portera le code `EV#EV03_2`. Ce type de code servira par la suite à établir la correspondance entre une racine et sa classe de postpositions.

Notons ici que le coréen se distingue des autres langues par l'emploi d'un double

système d'écriture. Les mots sont constitués de caractères syllabiques Hangul qui sont des représentations de suites de lettres Jamo. Ainsi, le caractère Hangul 가 correspond aux deux lettres Jamo ᄀ et ᅡ. Le problème est que les variations subies par les racines ne correspondent pas toujours à des caractères Hangul, comme c'est le cas dans l'exemple de la figure 1. Il a donc été nécessaire de gérer le passage d'un système d'écriture à l'autre. Par ailleurs, le coréen autorise l'emploi de certains caractères chinois en remplacement de caractères Hangul. Ce phénomène a été géré par l'établissement d'une liste des correspondances autorisées dont voici un court extrait:

諫간
間간
刊갈
喝갈

Grâce à cette liste, le linguiste n'a pas à se préoccuper de ce type de variantes et peut se contenter de tenir à jour un dictionnaire des formes écrites en coréen, le système de consultation de dictionnaire se chargeant d'établir automatiquement les correspondances avec les caractères chinois.

Au total, dans le cas des verbes, les classes de variantes de racines sont au nombre de 64. À l'issue de la phase de génération des variantes des racines, on obtient un dictionnaire de racines au format DELAF que l'on transforme en automate, aussi bien pour le compresser que pour en accélérer la consultation. Pour des raisons d'efficacité, les entrées sont converties sous forme de suites de lettres Jamo avant d'être compressées sous forme d'automate. En effet, la complexité de la recherche d'un mot dans un automate est en *taille alphabet* × *longueur du mot*. Or, la taille de l'alphabet Jamo est inférieure à 30 lettres alors que le nombre de caractères Hangul est supérieur à 11000.

2.2 Description des classes de postpositions

La combinatoire des postpositions est complexe, mais comporte néanmoins de nombreuses régularités. Pour cette raison, les classes de postpositions sont décrites au moyen de grammaires modulaires pouvant s'appeler les unes les autres, afin de factoriser les descriptions redondantes. Chaque classe est caractérisée par sa grammaire principale dont le nom correspond à l'un des codes produits à l'étape de génération des variantes de racines. Les grammaires de postpositions associent des étiquettes morpho-syntaxiques à des séquences constituées de caractères Hangul et/ou Jamo.

La figure 2 montre la grammaire des postpositions EV/EV03_2. Elle sera mise en correspondance avec les racines portant le code `EV#EV03_2`[1]. La figure 3 montre le sous-graphe SUG3_2 appelé depuis cette grammaire. On peut y voir que chaque morphème est associé à un étiquetage morpho-syntaxique.

1 Le dièse remplace dans les graphes le caractère / qui a déjà une utilisation particulière.

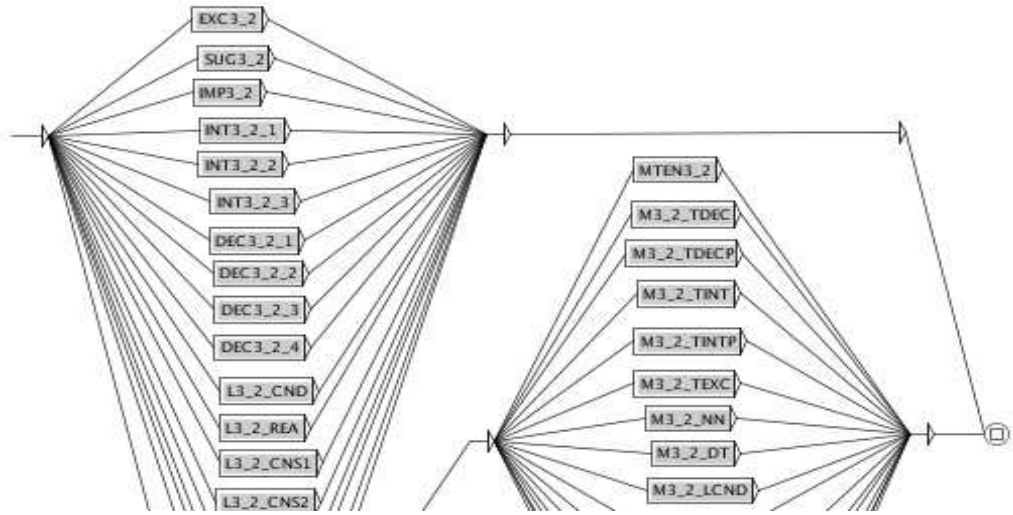

Figure 2: extrait du graphe de postpositions EV/EV03_2

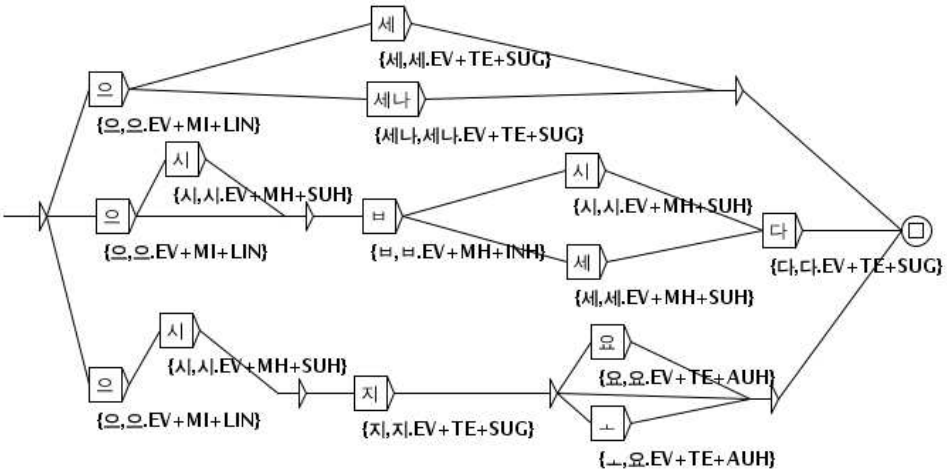

Figure 3: graphe de postpositions SUG3_2

L'ensemble des grammaires de postpositions est constitué de 2728 graphes.

2.3 Graphe dictionnaire

La mise en correspondance des racines avec leurs classes de postpositions se fait au moyen d'un graphe comme celui de la figure 4. Les symboles < et > qui entourent le contenu du graphe indiquent qu'il s'agit d'un graphe destiné à être appliqué caractère par caractère au texte que l'on souhaite analyser. Le symbole `<AS>` indique que l'on veut reconnaître une racine en consultant le dictionnaire de racines que l'on a construit précédemment. Lorsqu'on a reconnu une racine, les lignes comme `$AS.EQ=EA#EA23_2$` jouent le rôle de tests pour savoir quelle branche va ensuite être explorée. Ainsi, si la racine contient le code `EA#EA23_2`, on explorera ensuite la grammaire de postpositions `EA/EA23_2`[2] pour finir d'analyser la séquence de caractères trouvée dans le texte. Pour chaque chemin de la grammaire de

2 Dans un nom de sous-graphe, le caractère : remplace le caractère slash, pour la même raison que dans la note précédente.

postpositions qui permet d'atteindre la fin du mot du texte que l'on est en train d'analyser, on produira une analyse qui sera constituée de la racine reconnue ainsi que de la suite de postpositions construite par concaténation lors de l'exploration du chemin de la grammaire de postpositions.

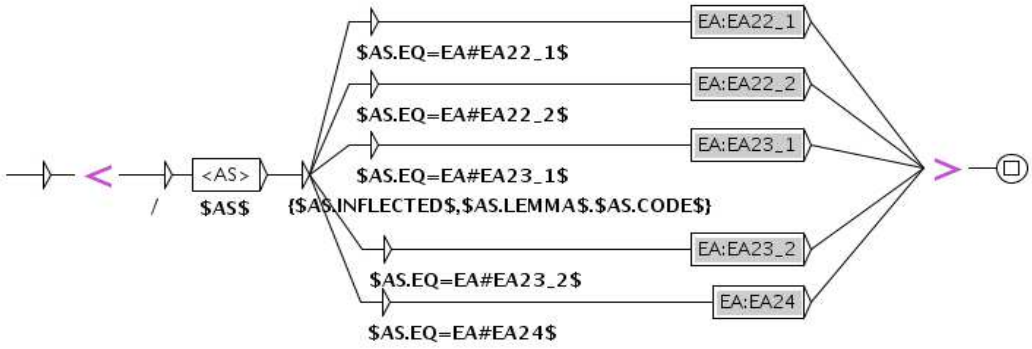

Figure 4: extrait du graphe dictionnaire des mots simples du coréen

Ce graphe est destiné à être appliqué au texte que l'on souhaite analyser par le programme de consultation de dictionnaire intégré à Unitex. L'analyse morphologique du coréen est ainsi ramenée à un classique problème de pattern matching. Le résultat de cette opération est un fichier listant pour chaque séquence reconnue, ses coordonnées dans le texte ainsi que la séquence de morphèmes étiquetés qui la compose. Ce fichier est ensuite utilisé pour construire pour chaque phrase du texte un automate décrivant toute la combinatoire des étiquettes morpho-syntaxiques reconnues, comme celui présenté sur la figure 5, dans lequel les transitions en pointillés entre deux boîtes signalent que les deux morphèmes représentés par ces boîtes appartiennent à un même mot typographique (Eojol).

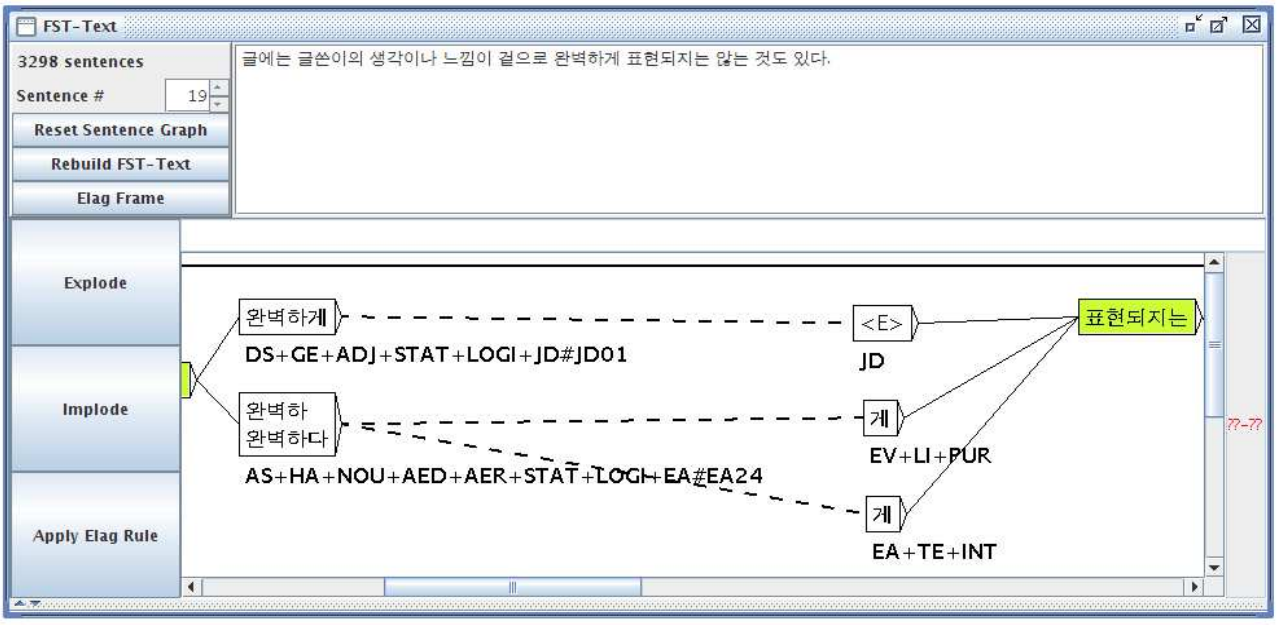

Figure 5: extrait d'un automate de phrase

Contrairement aux analyseurs du coréen existant comme Geuljabi (www.sejong.or.kr), ce système offre toutes les analyses possibles sous forme de parcours possibles dans les automates de phrase. Cette approche qui privilégie le rappel permet de ne

pas bloquer des analyses ultérieures en commettant des erreurs d'étiquetage tôt dans la chaîne de traitement d'un texte.

3. Performances

L'application du dictionnaire complet sur un texte codé en UTF-16LE de 275 Ko, contenant 3298 phrases, prend 4,5 secondes sur un PC Core 2 Duo sous Ubuntu à 2,4 Ghz et produit 48687 analyses pour 15881 séquences reconnues. Une fois cette étape terminée, la construction des automates de phrase prend 3 secondes. Sur un texte de 10 Mo contenant 365000 phrases, l'application prend 1m37s pour 1780546 analyses correspondant à 736308 séquences reconnues. La construction des automates de phrases prend 4m8s. Malgré l'augmentation de complexité par rapport aux mécanismes utilisés pour les langues non agglutinantes, ces temps de traitement sont tout à faits acceptables pour des besoins applicatifs. Il nous manque encore une évaluation humaine complète du dictionnaire produit pour le coréen pour vérifier qu'il ne contient pas d'erreurs, mais la mise en œuvre complète du système sur des données à grande échelle a d'ores et déjà permis de faire la preuve de sa viabilité.

4. Conclusion

Le modèle de système de dictionnaire que nous avons construit pour le coréen offre plusieurs avantages. Il est simple à utiliser, car les différentes données sont toutes éditables aisément sous une forme graphique, ce qui est particulièrement utile pour décrire la combinatoire des postpositions. Ainsi, toute la complexité a été transféré des données vers les programmes chargés de les manipuler, ce qui fait que les utilisateurs linguistes n'ont besoin d'aucune compétence particulière pour maîtriser un formalisme de description complexe. De plus, la technique mise en oeuvre peut être directement réutilisée pour les autres langues agglutinantes. Cela a notamment déjà été le cas pour gérer des cas d'agglutination en arabe (Neme 2011). Cette technique a également été étendue avec succès au traitement des mots composés du coréen, pour lequel la procédure de flexion reprend une partie de la flexion des mots simples. Enfin, tous les mécanismes utilisés sont pleinement opérationnels et diffusés dans le logiciel libre Unitex.

Références

Beesley, K., Karttunen, L. 2003. *Finite State Morphology*. CSLI Publications.

Courtois, B. 1990. Un système de dictionnaires électroniques pour les mots simples du français, *Langue Française* 87, Paris: Larousse, pp. 11-22

Gross, M. 1989. La construction de dictionnaires électroniques. *Annales des Télécommunications*, tome 44, nø 1-2, pp. 4-19, Issy-les-Moulineaux/ Lannion: CNET.

Han Ch. H., Palmer, M. 2005. A Morphological Tagger for Korean: Statistical Tagging Combined with Corpus-based Morphological Rule Application. *MT journal*.

Huh, H.-G. 2005. *Délimitation et étiquetage des morphèmes en coréen par ressources linguistiques.* Thèse de doctorat. Université de Marne-la-Vallée.

Koskenniemi, K. 1984. A general computational model for word-form recognition and production. In *Proceedings of the 10th international Conference on Computational Linguistics and 22nd Annual Meeting on Association For Computational Linguistics*

(Stanford, California, July 02 - 06, 1984). Annual Meeting of the ACL. Association for Computational Linguistics, Morristown, NJ, 178-181.

Lee, C.-Y. 1997. *La construction de lexiques de formes fléchies et l'analyse morphologique du coréen*. Thèse de doctorat. Université Paris 7.

Nam, J.-S. 2002. Construction of the Sub-modules of Korean Electronic Dictionary of Nouns DECO-N. *HUFS Dissertations* N-34. Hankuk University of Foreign Studies. Korea. 105-125.

Nam, J.-S. 2003. Some issues on the construction of the electronic lexicon of Korean adjectives. *Language Research* 39-1. Seoul National University. Korea. 205-241.

Nam, J.-S. 2007. *Inflection of Korean Verbs and Adjectives DECOP*. Parkleejung Publishing Company. Korea.

Neme, A. 2011. *A lexicon of Arabic verbs constructed on the basis of Semitic taxonomy and using finite-state transducers*. (accepted for WoLeR 2011)

Paumier, S. 2010. Unitex 2.1 User Manual. http://igm.univ-mlv.fr/~unitex

Revuz, D. 1991. *Dictionnaires et lexiques: méthodes et algorithmes*. Thèse de doctorat. Université Paris 7.

Roche, E., Schabès, Y. (eds.). 1997. *Finite-State Language Processing*. Cambridge, Mass./ London, MIT Press.

Silberztein, M. 1993. *Dictionnaires électroniques et analyse automatique de textes – le système INTEX*. Masson. Paris.

Silberztein, M. 1999. INTEX: a Finite State Transducer Toolbox. *Theoretical computer science*. Vol 231:1, pp. 33-46.